ISSN: 2690-909X

Research Article

Advances in Neurology and Neuroscience

Limitations of Generative AI for Numerical Aggregation of Apple Health Step Count Data: A Validation Study Using XML Records, Manual Calculation, and Health Auto Export

Zhenghua Li[1,2] and Kenichi Yamamura[1,2*]

[1]*Institute of Resource Development and Analysis, Kumamoto University, Kumamoto, Japan*

[2]*Transgenic Group, Inc., Fukuoka, Japan*

[*]Corresponding Author
Kenichi Yamamura, Institute of Resource Development and Analysis, Kumamoto University, Kumamoto, Japan.

Submitted: 2026, May 11; **Accepted:** 2026, Jun 05; **Published:** 2026, Jun 16

Citation: Li, Z., Yamamura, K. (2026). Limitations of Generative AI for Numerical Aggregation of Apple Health Step Count Data: A Validation Study Using XML Records, Manual Calculation, and Health Auto Export. *Adv Neur Sci, 9*(2), 01-07.

Abstract

Background: *Generative artificial intelligence (AI) tools are increasingly used to assist data handling and research workflows. However, their reliability for numerical aggregation of raw health data has not been fully validated.*
Objective: This study examined whether generative AI could accurately aggregate Apple Health step count data from original XML records and compared the results with manual calculation and Health Auto Export.

Methods: *Step count records were extracted from Apple Health XML data. The original data consisted of multiple measurement records per day. Daily step counts were calculated using four approaches: two independent aggregations using Apple Intelligence plus ChatGPT, Health Auto Export, and manual calculation in Excel from the original XML-derived records. Daily totals and weekly summaries were compared across methods.*

Results: *The two AI-based aggregations produced different daily and total step counts and did not agree with either Health Auto Export or manual calculation. In contrast, Health Auto Export and manual Excel calculation showed complete agreement for all daily step counts examined. The initial AI-based analysis also incorrectly treated the data as steps per measurement rather than steps per day.*

Conclusion: *Generative AI was unreliable for primary numerical aggregation of Apple Health step count data, even for simple summation tasks. Health Auto Export, validated against manual calculation, provided reproducible daily step count data suitable for subsequent analysis. Generative AI may be useful for research planning, interpretation, and manuscript preparation, but raw data extraction, numerical aggregation, and statistical calculation should be performed using reproducible computational tools.*

Keywords: Generative Artificial Intelligence, ChatGPT, Apple Health, Health Auto Export, Step Count, Wearable Devices, Data Validation, Numerical Aggregation

1. Introduction

Wearable devices and smartphones provide continuous or semi-continuous health-related data that may be useful for longitudinal assessment of disease progression, daily activity, and treatment response [1-6]. Apple Health data, including step count and gait-related metrics, can be exported as original XML records, and HealthKit provides a centralized framework for health and fitness data collected from iPhone, Apple Watch, and related applications [7,8]. Recent work has also used Apple HealthKit-derived digital phenotyping to identify pharmacological interference in Parkinsonian gait, highlighting the practical value of smartphone-derived health data in clinical research [1]. At the same time, generative artificial intelligence (AI) tools, including ChatGPT-based systems, are increasingly used in research workflows [2]. These tools are useful for summarizing information, organizing methods, drafting manuscripts, and interpreting results. However, their reliability for numerical data extraction and aggregation remains uncertain, particularly because recent studies have shown

that large language models can make arithmetic and numerical reasoning errors under certain conditions [3].

This issue is particularly important when dealing with health data. Even apparently simple tasks, such as summing step count records by date, require accurate identification of relevant records, correct grouping by day, and complete aggregation of all values. Errors at this stage directly affect downstream analyses and conclusions. The present study was prompted by discrepancies in initial step count analyses. The original Apple Health XML data consisted of measurement-level records rather than daily totals. Therefore, a validation analysis was performed to determine whether AI-based aggregation could accurately reproduce daily step counts. AI-based results were compared with Health Auto Export and manual calculation from XML-derived records.

2. Methods

2.1. Data Source

Step count data were obtained from Apple Health XML export files. The XML export did not provide one daily step-count value. Instead, each record represented a measurement-level event and included the record type, source name, device information, unit, creation date, start date, end date, and value. Because several records could occur within the same day, daily totals first had to be calculated by summing all values with the same calendar date before weekly mean steps/day values could be derived for analysis (Table 1). The structure of the original XML records and their conversion into an Excel-based measurement-level table are shown in Table 2. Briefly, the XML text was opened in a text editor, copied into Excel, and separated into columns. Nonessential fields were removed, leaving date, start time, end time, and step-count value, daily totals were then calculated from this measurement-level table.

Week Range	SUM	No. of raw records	Mean (Steps/measurement)
2024/01/01–01/07	33,969	211	160.99
2024/01/08–01/14	45,208	230	196.56
2024/01/15–01/21	27,126	180	150.70
2024/01/22–01/28	30,936	188	164.55

A. Correct aggregation of raw XML data

Week Range	SUM	No. of days	Mean (steps/day)
2024/01/01–01/07	33,992	7	4,856.00
2024/01/08–01/14	44,178	7	6,311.10
2024/01/15–01/21	28,722	7	4,103.10
2024/01/22–01/28	40,850	7	5,835.70
Note: Because step count varies from day to day, daily totals must first be calculated before weekly mean steps/day values are obtained			

B. Incorrect aggregation using measurement-level records

Table 1: Initial Aggregation Error Caused by Treating Measurement-level Records as Daily Data

Step 1. Export the XML file from iPhone.
Step 2. Open the XML file in TextEdit. Each step-count record is displayed as shown below:
<Record type="HKQuantityTypeIdentifierStepCount"
sourceName="iPhone 15 plus ky"
sourceVersion="17.1.2"
device="<<HKDevice: 0xc6a7fed00>, name:iPhone, manufacturer:Apple Inc., model:iPhone, hardware:iPhone15,5, software:17.1.2, creation date:2023-12-03 00:49:48 +0000>"
unit="count"
creationDate="2024-01-01 08:35:59 +0900"
startDate="2024-01-01 08:25:49 +0900"
endDate="2024-01-01 08:25:54 +0900"
value="11"/>
Step 3.
Copy the XML text into an Excel file and split the text into columns. The resulting table shows the data for each measurement record.

<Record	type=	HKQuantityTypeIdentifierStepCount	sourceName=	iPhone	15plus	ky	sourceVersion=	17.1.2	device=
<Record	type=	HKQuantityTypeIdentifierStepCount	sourceName=	iPhone	15plus	ky	sourceVersion=	17.1.2	device=
<Record	type=	HKQuantityTypeIdentifierStepCount	sourceName=	iPhone	15plus	ky	sourceVersion=	17.1.2	device=
<Record	type=	HKQuantityTypeIdentifierStepCount	sourceName=	iPhone	15plus	ky	sourceVersion=	17.1.2	device=
<Record	type=	HKQuantityTypeIdentifierStepCount	sourceName=	iPhone	15plus	ky	sourceVersion=	17.1.2	device=

<<HKDevice:	name:iPhone,	manufacturer:Apple	Inc.,	model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	manufacturer:Apple
<<HKDevice:	name:iPhone,	manufacturer:Apple	Inc.,	model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	manufacturer:Apple
<<HKDevice:	name:iPhone,	manufacturer:Apple	Inc.,	model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	manufacturer:Apple
<<HKDevice:	name:iPhone,	manufacturer:Apple	Inc.,	model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	manufacturer:Apple
<<HKDevice:	name:iPhone,	manufacturer:Apple	Inc.,	model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	manufacturer:Apple

Inc., zodel:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	date:2023-12-03	0:49:48	+0000>	unit=	count	creationDate=
Inc., model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	date:2023-12-03	0:49:48	+0000>	unit=	count	creationDate=
Inc., model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	date:2023-12-03	0:49:48	+0000>	unit=	count	creationDate=
Inc., model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	date:2023-12-03	0:49:48	+0000>	unit=	count	creationDate=
Inc., model:iPhone,	hardware:iPhone15,5,	software:17.1.2,	creation	date:2023-12-03	0:49:48	+0000>	unit=	count	creationDate=

0xc6a7fed00>,	startDate=	2024/1/1	8:25:49	900	endDate=	2024/1/1	8:25:54	900	value=	11	/>
0xc6a7fed00>,	startDate=	2024/1/1	8:58:05	900	endDate=	2024/1/1	8:58:08	900	value=	16	/>
0xc6a7fed00>,	startDate=	2024/1/1	9:13:03	900	endDate=	2024/1/1	9:13:06	900	value=	8	/>
0xc6a7fed00>,	startDate=	2024/1/1	9:26:37	900	endDate=	2024/1/1	9:26:52	900	value=	35	/>
0xc6a7fed00>,	startDate=	2024/1/1	10:17:02	900	endDate=	2024/1/1	10:22:03	900	value=	37	/>

Step 4.

After retaining only the required columns, the measurement-level table shown on the left is obtained (January 1 only). This table is the basis for aggregation. Daily totals are then calculated for each date, as shown on the right.

Start Date	Start Time	End Time	Value
2024/1/1	8:25:49	8:25:54	11
2024/1/1	8:58:05	8:58:08	16
2024/1/1	9:13:03	9:13:06	8
2024/1/1	9:26:37	9:26:52	35
2024/1/1	10:17:02	10:22:03	37
2024/1/1	10:33:19	10:34:33	42
2024/1/1	11:20:54	11:25:45	34
2024/1/1	12:06:29	12:13:10	39
2024/1/1	12:43:41	12:47:16	32
2024/1/1	13:38:37	13:42:55	53
2024/1/1	15:31:32	15:40:08	83
2024/1/1	18:25:00	18:27:49	36
2024/1/1	18:59:50	19:03:04	39
2024/1/1	19:37:25	19:41:22	18
2024/1/1	20:13:10	20:17:02	24
2024/1/1	20:30:51	20:33:42	64
2024/1/1	21:21:30	21:23:55	40
2024/1/1	22:07:35	22:13:37	40

Date	SUM	Date	SUM
2024/1/1	651	2024/1/15	5507
2024/1/2	9330	2024/1/16	7658
2024/1/3	528	2024/1/17	2976
2024/1/4	8412	2024/1/18	564
2024/1/5	1882	2024/1/19	7595
2024/1/6	11333	2024/1/20	2235
2024/1/7	1833	2024/1/21	591
2024/1/8	8528	2024/1/22	4871
2024/1/9	5827	2024/1/23	459
2024/1/10	7854	2024/1/24	5603
2024/1/11	5747	2024/1/25	3899
2024/1/12	6300	2024/1/26	7417
2024/1/13	3739	2024/1/27	2192
2024/1/14	7213	2024/1/28	6495

Table 2: Structure of Apple Health XML Records and Conversion into a Measurement-level Excel Table

2.2. Aggregation Methods

Daily step count values were calculated using four approaches (Tables 3 and 4):

- First aggregation using Apple Intelligence plus ChatGPT
- Second independent aggregation using Apple Intelligence plus ChatGPT
- Health Auto Export daily step count data. Health Auto Export is a third-party application that exports Apple Health data into structured formats, including daily summary data.
- Manual calculation in Excel from the XML-derived measurement-level records. For the manual calculation, all measurement-level step count values belonging to the same calendar date were summed to obtain the daily step count.

Date	First AI aggregation	Second AI aggregation	Health Auto Export	Manual Excel calculation
2024/1/1	628	618	651	651
2024/1/2	8726	9176	9330	9330
2024/1/3	528	494	528	528
2024/1/4	8220	7802	8412	8412
2024/1/5	1828	1778	1882	1882
2024/1/6	11141	9760	11333	11333
2024/1/7	1921	1713	1833	1833
2024/1/8	8412	7360	8528	8528
2024/1/9	7263	5801	5827	5827
2024/1/10	6478	7253	7854	7854
2024/1/11	5589	5921	5747	5747
2024/1/12	5969	6240	6300	6300
2024/1/13	4859	3970	3739	3739
2024/1/14	6610	5885	7213	7213
2024/1/15	5897	5374	5507	5507
2024/1/16	7470	6333	7658	7658
2024/1/17	4087	2804	2976	2976
2024/1/18	688	554	564	564
2024/1/19	7823	6890	7595	7595
2024/1/20	3065	2120	2235	2235
2024/1/21	692	563	591	591
2024/1/22	5381	5012	4871	4871
2024/1/23	520	473	459	459
2024/1/24	7773	5355	5603	5603
2024/1/25	6352	4267	3899	3899
2024/1/26	9500	7341	7417	7417
2024/1/27	2737	2122	2192	2192
2024/1/28	8587	6344	6495	6495
Total	148744	129323	137239	137239
The two Apple Intelligence plus ChatGPT aggregations produced inconsistent results and did not match the Health Auto Export or manual Excel results. Health Auto Export and manual Excel calculation matched exactly.				

Table 3: Comparison of Daily Step Count Totals by Aggregation Method.

Week Range	SUM	No. of days	Mean (steps/day)
2024/01/01–01/07	33,992	7	4,856
2024/01/08–01/14	44,178	7	6,311
2024/01/15–01/21	28,722	7	4,103
2024/01/22–01/28	40,850	7	5,835

Week Range	SUM	No. of days	Mean (steps/day)
2024/01/01–01/07	33,492	7	4,784
2024/01/08–01/14	45,180	7	6,454
2024/01/15–01/21	28,819	7	4,117
2024/01/22–01/28	40,850	7	5,835

A. Weekly summaries obtained from two independent aggregations using Apple Intelligence plus ChatGPT

Week Range	SUM	No. of days	Mean (steps/day)
2024/01/01–01/07	33,969	7	4,852.71
2024/01/08–01/14	45,208	7	6,458.29
2024/01/15–01/21	27,126	7	3,875.14
2024/01/22–01/28	30,936	7	4,419.43
(1) Manual daily aggregation followed by manual weekly calculation in Excel Week Range			

Week Range	SUM	No. of days	Mean (steps/day)
2024/01/01–01/07	34,969	7	4,995.57
2024/01/08–01/14	45,208	7	6,458.29
2024/01/15–01/21	22,466	7	3,209.43
2024/01/22–01/28	35,936	7	5,133.71
(2) Manual daily aggregation followed by AI-assisted weekly calculation			

B. Comparison of weekly summaries generated from manually curated data

C. Result (Note):

The two Apple Intelligence plus ChatGPT aggregations produced different weekly summaries, and both differed from the Health Auto Export and manual Excel results. In contrast, Health Auto Export and manual Excel calculation matched exactly.

2.3. Weekly Summary

Daily step counts were further summarized into weekly periods. Step count was treated as a descriptive measure of daily walking activity. Because daily step count is strongly influenced by activity opportunity, environmental conditions, rehabilitation days, and daily schedule, standard deviation and statistical significance testing were not applied to step count in the final analysis.

2.4. Validation

The daily values obtained by each method were compared. Agreement between methods was evaluated by direct comparison of daily totals and overall totals.

3. Results

3.1. Original Data Structure

The original XML data were not daily summary data. They consisted of multiple measurement-level records for each day (Table 2). Therefore, direct division of weekly total step count by the number of measurement records produced steps per measurement, not steps per day. This was identified as the main methodological error in the initial analysis.

3.2. Discrepancy in AI-Based Aggregation

The first and second AI-based aggregations produced different daily values (Table 3). They also differed from both Health Auto Export and manual Excel calculation. This indicated that the AI-based aggregation was not reproducible. For example, daily step counts calculated by AI differed between the first and second attempts, and both differed from the validated values. The total step count for the examined period also differed across AI attempts, whereas Health Auto Export and manual calculation showed identical totals (Table 3).

3.3. Agreement Between Health Auto Export and Manual Calculation

Health Auto Export and manual Excel calculation from the XML-derived records showed complete agreement for all examined dates. This confirmed that Health Auto Export accurately reproduced the daily step count values from the original XML records. Therefore, Health Auto Export daily data were considered reliable for subsequent weekly aggregation and descriptive analysis. These findings are summarized in Tables 1-4, which show that the two AI-based aggregations differed from each other and from the validated values, whereas Health Auto Export and manual calculation matched exactly.

4. Discussion

This study demonstrates a practical limitation of generative AI in numerical aggregation of health data. Although step count aggregation appears to be a simple summation task, AI-based aggregation produced inconsistent and inaccurate results. In contrast, Health Auto Export and manual Excel calculation showed complete agreement. The main reason is that generative AI systems are not designed as deterministic calculation engines. They process text and numbers as language-like tokens and generate plausible outputs based on patterns. When large numbers of records are provided, the model may omit records, duplicate entries, misclassify dates, or produce internally inconsistent results. The resulting values may appear reasonable, but they are not guaranteed to be complete or reproducible. In this study, the problem was not limited to arithmetic. The more important issue was preprocessing. Correct aggregation required identifying all records belonging to each date, extracting the correct value field, summing all values for that date, and distinguishing measurement-level data from daily summary data. An error in any of these steps changes the final daily step count.

The initial analysis also revealed a conceptual problem. Because the original XML data were measurement-level records, dividing weekly total step count by the number of measurement records produced steps per measurement. This is not an appropriate representation of daily physical activity. Step count should be calculated as total steps per day and then summarized descriptively over weekly or monthly periods. These findings do not imply that generative AI is useless in research. Rather, they clarify the appropriate division of labor. Generative AI is highly useful for organizing research questions, identifying methodological problems, drafting manuscript sections, improving clarity of scientific writing, and discussing interpretation. It can also help researchers recognize when the analysis unit is inappropriate, as in the distinction between steps per measurement and steps per day. This interpretation is consistent with current publication guidance, which emphasizes human accountability, transparency, and disclosure when AI-assisted tools are used in scholarly work [4-6].

However, generative AI should not be used as the primary tool for raw data extraction, numerical aggregation, statistical calculation, or final table generation unless the results are independently verified. For these tasks, deterministic and reproducible tools such as Excel, Numbers, R, Python, or validated export software should be used. This distinction is particularly important for medical and health-related research. Numerical errors in primary data aggregation can lead to incorrect conclusions about disease progression, treatment effects, or functional decline. Because AI-generated numerical outputs may look plausible, such errors may not be detected unless compared against reproducible calculation methods. Reporting frameworks for AI-related clinical research also emphasize transparent description of the AI system, its inputs and outputs, human-AI interaction, and error cases [9]. The present validation supports the use of Health Auto Export for daily step count extraction, at least for the examined dataset, because it exactly matched manual calculation from the original XML records. Once validated, such software can reduce the workload of handling large-scale longitudinal health data while maintaining numerical reliability.

Step count itself should also be interpreted cautiously. Unlike gait metrics such as walking speed, step length, double support percentage, or walking asymmetry, step count reflects daily activity volume rather than gait quality. It is influenced by external factors such as weather, appointments, rehabilitation sessions, use of mobility aids, and opportunity to walk. Therefore, in the present analysis, step count was treated as a descriptive activity measure rather than a variable for statistical significance testing. Overall, the present study suggests that generative AI should be used as an intellectual and editorial assistant, not as an unverified numerical data processor. The most efficient and reliable workflow is to use reproducible tools for data extraction and calculation, and to use AI for interpretation, methodological refinement, and manuscript preparation. By combining a concrete error example with a practical validation workflow, the present report may be particularly useful for AI users in health-related research who need to decide which parts of the workflow can be safely assisted by generative AI and which require deterministic validation.

5. Conclusion

Generative AI produced inconsistent and inaccurate daily step count values when used to aggregate Apple Health XML records. In contrast, Health Auto Export showed complete agreement with manual Excel calculation and was therefore suitable for subsequent analysis. The appropriate role of generative AI in this context is not primary numerical aggregation, but support for study design, interpretation, and scientific writing. For health data analysis, raw data extraction and numerical calculation should be performed using reproducible and independently verifiable methods. Based on these findings, we propose a practical workflow for AI-assisted analysis of wearable-derived health data (Table 5). The key principle is that numerical extraction, aggregation, and statistical calculation should be performed using reproducible tools, whereas generative AI should be used for study planning, interpretation, and manuscript preparation.

Step	Recommended practice	Rationale
1	Export raw or daily health data using a reproducible tool	Ensures that the same input data can be reanalyzed
2	Validate a subset against original records or manual calculation	Detects extraction or aggregation errors
3	Define the correct analytical unit before analysis	Prevents confusion between measurement-level and daily-level data
4	Use deterministic tools for numerical aggregation and statistics	Avoids non-reproducible AI-generated numerical errors
5	Treat generative AI as a support tool for interpretation and writing	Uses AI where it is strongest while avoiding unsupported numerical processing
6	Disclose AI use transparently in the manuscript	Maintains accountability and publication transparency

Table 5: Recommended Workflow for AI-Assisted Analysis of Wearable-Derived Health Data

Declaration of Generative AI Use: Generative AI was used to assist with drafting and language editing of this manuscript. Numerical data extraction, validation, and calculation were performed using Health Auto Export and manual Excel calculation, not by generative AI. Final interpretation and responsibility for the manuscript remain with the authors.

Author Contributions: K.Y. conceived the study, interpreted the data, and wrote the manuscript. Z.L. curated the data and contributed to formal analysis. Both authors reviewed and approved the final manuscript.

Funding: This research received no external funding.

Competing Interests: The authors declare no competing interests.

Ethics Approval and Consent to Participate: This study analyzed anonymized self-collected gait data from a single individual. The participant provided informed consent for the use and publication of these data. According to institutional guidelines, analysis of de-identified self-tracked data did not require IRB approval.
Consent for Publication: The participant provided consent for publication.

Data Availability: Data are available from the corresponding author upon reasonable request.

References

1. Li, Z., & Yamamura, K. (2026). Objective Identification of Pharmacological Interference in Parkinsonian Gait Using Continuous Digital Phenotyping.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33,* 1877-1901.
3. Shrestha, S., Kim, M., & Ross, K. (2025). Mathematical reasoning in large language models: Assessing logical and arithmetic errors across wide numerical ranges. *arXiv preprint arXiv:2502.08680.*
4. Zielinski, C., Winker, M., Aggarwal, R., Ferris, L., Heinemann, M., Lapeña, J. F., ... & Habibzadeh, F. (2023). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications revised May 31, 2023. *Philippine Journal of Otolaryngology Head and Neck Surgery, 38*(1), 7-7.
5. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Use of Artificial Intelligence by Authors.
6. Nature Portfolio. Artificial Intelligence editorial policy.
7. Apple Inc. Share your health and fitness data in XML format. Apple Support.
8. Apple Inc. HealthKit. Apple Developer Documentation.
9. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., ... & Yau, C. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health, 2*(10), e537-e548.